5. CONCEPT AND DESIGN OF RETROSPECTIVE STUDIES IN NUTRITIONAL EPIDEMIOLOGY

Wieslaw A. Jedrychowski, Umberto Maugeri

Epidemiology is a science about the spread of diseases in human populations, their onset (incidence) and prevalence, disabilities, or deaths with the main goal to identify causes and risk factors responsible for their occurrence. Although, epidemiology shares its interest in human diseases with other medical disciplines but their primary concern is focused on occurrence and spread of a disease in a population and it is concerned with human beings, as members of a community living in their integral environment. In essence, epidemiology describes the natural history of diseases, with their complicated interrelations between the environment, life style and the genetically determined susceptibility to diseases. Particular and ultimate goal of epidemiology is to institute preventive measures against diseases and strengthen public health.

Nutritional epidemiology can be defined as the study of the nutritional determinants of disease in human populations. It has wide-ranging goals, but the most significant tasks are monitoring the dietary habits, food consumption, nutrient intake and nutritional status of populations. This information combined with the health status of a given population may help to open the way to the new knowledge about nutrition-related diseases.

Since long the role of dietary habits has been considered as a key factor influencing public health. Already Hippokrates (400 years BC) tried to persuade his students about the importance of healthy dietary habits and this point was well documented in one of his papers "...one should consider most attentively the water which the inhabitants use, whether it is marshy and soft, or hard and running from elevated and rocky situations, and then if saltish and unfit for cooking; and the ground, whether it be naked and deficient in water, or wooded and well watered, and whether it lies in a hollow, confined situations, or is elevated and cold; and the mode in which the inhabitants live, and what are their pursuits, whether they are fond of drinking and eating to excess, and given to indolence, or are fond of exercise and labor...".

Although in the past epidemiologists focused their attention on epidemics of infectious diseases, in the early decades of the 20th century nutritional epidemiology gained ground being engaged in a series of many studies aiming at understanding the nature of nutrition-related diseases. The investigations of scurvy or pellagra are spectacular and frequently mentioned examples of epidemiologic reasoning illustrating the power of epidemiologic observations and design strategy. Another excellent example are epidemiologic studies, which showed the beneficial effect of supplementation of folic acid in early

periods of pregnancy on the risk of delivering a child with a neural tube birth defect. Up to now, the mechanism of action of folic acid is not fully understood, but public health authorities already embarked on taking very successful preventive action based on this new knowledge.

Nutritional epidemiology is very important research field but not easy type of population studies. The importance of these studies comes from their significance for many present-day health problems, such as cardiovascular diseases (heart attacks, stroke), cancer, diabetes, congenital malformations, and may others. All these diseases are the scope of nutritional epidemiology studies and some of the findings have already been put into preventive practice. A major difficulty of nutritional epidemiology results from the very complex nature of immense variety of dietary factors potentially involved in the etiology of diseases. The foods that people eat are complex mixtures of various food products and compounds, which may be apparently similar. Moreover, people who eat more of one type of food may eat less of other types of foods, and it is extremely difficult to disentangle a complex set of inter-correlations among various dietary components. Furthermore, eating habits may be correlated with other potential factors involved in the disease web of causation such as socioeconomic status, life style characteristics or genetic traits. In addition, the various practices in food preparation are also important, and eating patterns within individuals often undergo transformation over years and interviewed people usually do not remember when and how quickly their dietary habits changed.

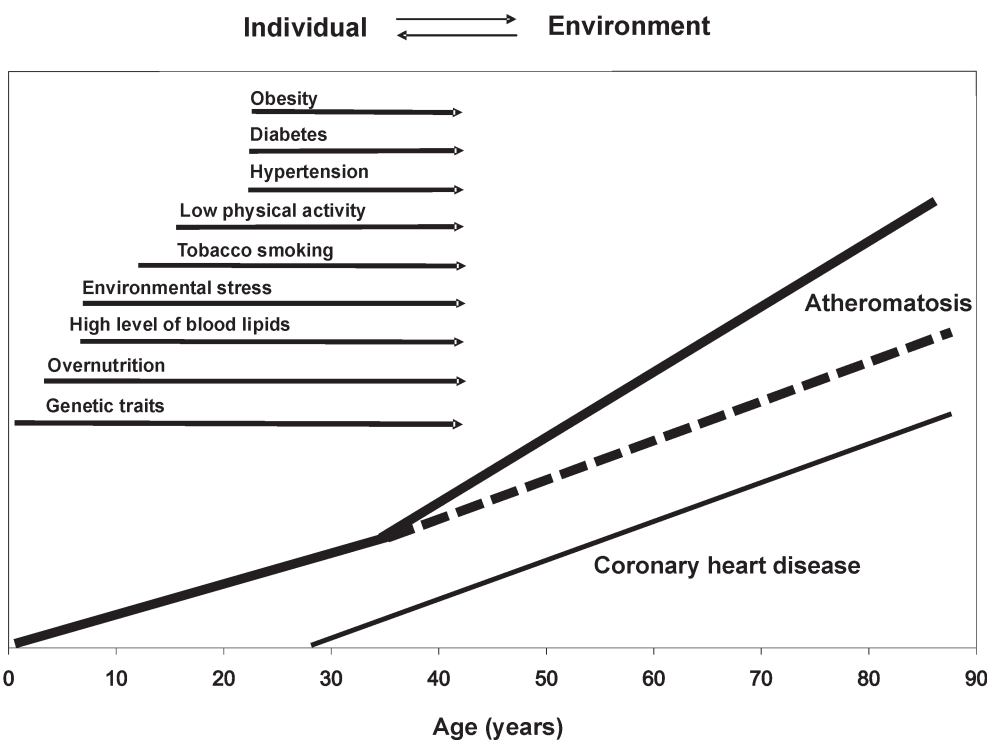


Figure 5.1. Nutrition is only one of the risk factors implemented in the occurrence of chronic diseases

It must be remembered that nature of chronic diseases is very complex in a sense that they have multiple causes and risk factors, which have usually very long and various latent periods and occur with relatively low frequencies, even among people with high exposure level (Figure 5.1). The most important weakness of nutritional epidemiology is the potential source of exposure bias pertinent to the measurement of dietary habits. Bias, defined as systematic error, producing an over- or underestimation of the exposure subsequently weakens strength of an association between a given exposure and an outcome. Another weakness of epidemiology is certainly the considerable difficulty in determining whether observed associations are causal. This is the crucial issue because the non-causal association between a given dietetic factor and a disease is irrelevant for the preventive practice.

Basic study designs in nutritional epidemiology

In order to establish sound scientific basis for the prevention of diseases and introducing effective health care programs epidemiologists use various methodological tools. On the whole, all research methods may be grouped into experimental and non-experimental design. Figure 5.2 presents the algorithm for classification of epidemiologic studies.

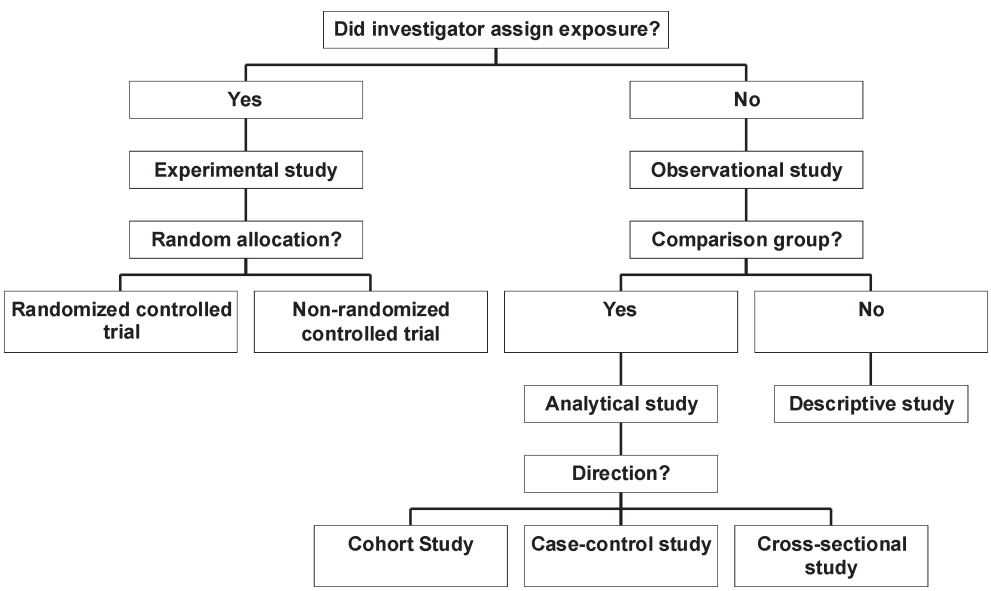


Figure 5.2. Algorithm for classification of types of epidemiologic research

Epidemiologic experimental design differ significantly from observational study type. Experimental investigations involve intentional intervention on the part of the investigator who is assessing the health context of given factors (e.g., specific vitamin supplementation), and allocating the subjects (patients) to the experimental groups

(supplemented vs. non-supplemented). Such opportunities are not existing in the non-experimental approach (observational design), where observations on effects of vitamin supplementation are made on the naturally established population groups.

Epidemiologic observational (non-experimental) studies are divided into two broad categories: *descriptive and analytic*. Descriptive epidemiology is the simple analysis of the distribution of diseases, exposures or other factors of interest within a given population in terms of person, place and time. Analytic epidemiology is the more precise study of the determinants of diseases, which requires appropriate control group.

At the basic level descriptive studies measure the occurrence (frequency) of a disease in a population in respect to demographic characteristics of the given population groups, which may be related to concrete type environment. Proxy measure of environment in descriptive studies is the occurrence of diseases in various population subgroups; e.g., those living in different geographic areas, climatic conditions, in various countries, or urbanized against rural regions. Dealing with microenvironment, one usually measures impact of occupation, incomes, residency, and living standards of subpopulations under study. The value of the descriptive study results from the fact that information collected in the course of descriptive study may lead to revealing a clue for possible causal relationship between specific dietary habits in various population groups and the occurrence of disease.

Analytic design (etiologic observational studies)

While the descriptive studies are focused on establishing preliminary associations between the incidence/prevalence of a disease and its potential causal factors of dietary origin, the etiologic or analytic studies are concerned with examining and interpretation of earlier observed facts in terms of cause-effect relationship. Etiologic relationship between disease and putative agent confirmed in the course of epidemiologic investigation has a great theoretical and practical meaning for the preventive action.

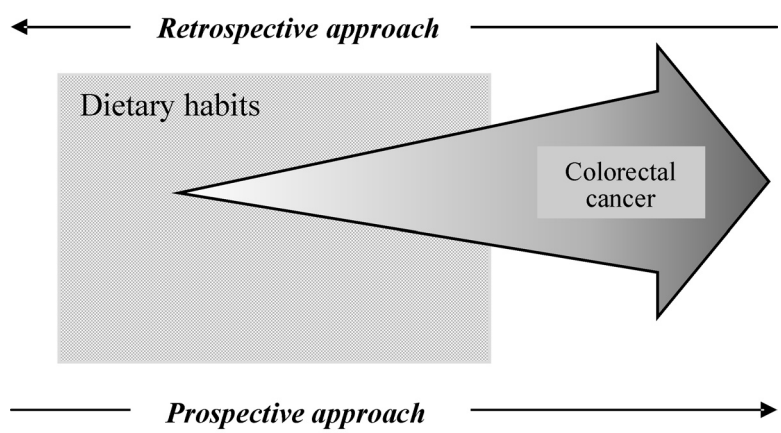


Figure 5.3. Direction of inquiry in prospective and retrospective epidemiologic studies

The way of collecting data for establishing etiological clues depends on whether the starting point in research is a disease itself or particular environment (dietary factors). In the so called **retrospective** or **case-control studies,** starting points are specified cases of disease, etiology of which is to be investigated by retrieving from the past the dietary data possibly having importance for the occurrence of disease (Figure 5.3). On the contrary, the **prospective studies** or **longitudinal** or **cohort studies** concentrate on a specific dietary exposure(s) and look forward for health outcomes possibly related causally to this exposure (cohort approach).

Study design of case-control (case-referent) studies

In a case-control study (retrospective approach), investigator first identifies people who have a disease **(cases**) and otherwise similar people who do not have it (**controls**) and compare their exposures to factors that may have influenced the disease risk in the past. Ideally, the cases and controls should be selected from the same population source and they should be representative of that population. The study is retrospective in the understanding that the spotlight of the study is on exposures that occurred in the past and on the ways in which these exposures may have affected an individual's actual health status of the groups under study.

Selection of cases

Selecting proper groups for the study is the crucial issue in the case-control study that poses some difficult problems. First, the problem of nosologic homogeneity of cases under study is of key importance, as it is more easy to deal with a single cause-effect chain, than with various several complicated etiologic issues considered at the same time. It is not enough to specify a given disease entity, but it is necessary to give full details of clinical stage of disease and specify diagnostic criteria (clinical/histological etc.) used for its identification. Accurate definition of the clinical stage will help to create homogeneous group of cases and minimize possible selection bias. This may seem contradictory to the requirement, that cases and controls should be representative of the target population. These are, however, two different issues, and precise definition of cases is not contradicting the representativeness of the groups to be chosen. Exaggerated pursuit in seeking for the perfect representative samples of cases often does not increase the precision of the study. Though representative samples help to extrapolate the results to target population, this cannot be done at the cost of the study precision. When we consider, a whole range of various clinical stages of a disease in a population, it may turn out that impact of a given exposure varies across the stages of a disease or population group.

The pursuit for perfect representative samples of cases comes from a belief that this would prevent biased sample selection. This may also be misleading, since validity of the case-control study predominantly depends on the most precise definition of a "case." Accurately formulated definition of cases grants precision of the estimated associations among the variables under study. Samples' representativeness is losing its significance unless the criteria of their selection are precisely defined.

If possible, the cases should always be drawn from the series of new cases clinically confirmed (incident cases), because cases diagnosed in the past exclude subjects who had a disease, but died, recovered, or changed the area of residence. It is difficult to dispute about the most appropriate source of cases – whether it is better to "retrieve" them from the population of hospitalized patients, case registries, or rather from the general population (outside of hospital). Generally, cases should be selected from hospitals, but only when a large proportion of patients is hospitalized due to the given disease. In other circumstances, cases drawn from the hospital population may differ too much from persons chosen outside of hospital, and this makes extrapolation of the results questionable.

Theoretically, it is not necessary to treat as cases all subjects with a disease at a given moment in a specified population source. A good source of cases may be patients taken from only one hospital, or even those receiving treatment from only one general practitioner. One issue is always crucial that source population for cases be precisely defined, because otherwise, it will be impossible to identify appropriate control group.

Population registers of cancer are valuable sources of cases and can usually provide referents as well. The registry may incorporate all cases of a particular disease such as cancer, poisonings, or malformation. Cases might also be drawn from a non- formal registry based on records collected for other purposes, such as hospital admission register, insurance claims, or disability pensions.

Selection of controls

As in selection of cases, precise definition of the control group is of vital importance for the results of the case-control study. Theoretically, control group should consist of persons from the source population, who have the same personal chance to become cases exactly in the same time as persons with a disease. Control groups should be drawn from the same target population as controls because the controls are supposed to be a reference level of the exposure occurring in the population at large. It is important to follow the rule, that controls must be drawn simultaneously with the cases. If both cases and controls have equal chance of hospital admission, then estimation of the risk of a disease based on the hospital samples is reliable.

There are two alternative approaches for selecting controls in case-control studies involving incident cases:

1. Density Sampling – one or more controls are selected for each case at the time of case detection – i.e., matching on time.

2. Cumulative Sampling – all controls are selected at the end of the observation period during which the cases are identified.

Density sampling is generally preferable when the observation period is long, especially if the frequency of exposure changes over time. Although control groups are often drawn from the hospitalized patients but they may be chosen from the open population rather than of neighbors, colleagues or relatives of cases. Random selection of the controls from outside of hospital is worthwhile, especially when cases have been also drawn from the open population. This is the most advisable method of ensuring the high level of comparability between groups, making possible the extrapolation of the conclusions. However, drawing cases and controls from outside of hospital is expensive, time-

consuming and more difficult in terms of co-operation with subjects recruited from the outside-hospital population.

There are some discrepancies in the opinions about the selection of controls in terms of diagnostic procedures. Some authors claim that controls should undergo exactly the same diagnostic procedures as their counterparts from the case group. This looks very reasonable, but hard to fulfill because of ethical considerations. It might happen that the same diagnostic management of controls could qualify some controls, say 1–2% as cases, because the disease has been latent. This, however, cannot bear negative implications on the study validity.

Selection of controls out of the persons, in whom diagnostic procedures excluded the disease in question, is not a perfect solution either. The same dietary factors (exposure) may once induce a disease A, and another time a disease B at other constellation of co-existing factors. Let us imagine a situation, where cases of peptic ulcer chosen from hospital, were matched to controls recruited from patients with chronic bronchitis who were also hospitalized in the same hospital. We may assume that both groups were subjected to the same diagnostic procedures (for example X-ray). Although we may incidentally find the association between peptic ulcer and cigarette smoking – the results of the study will definitively be biased. As cigarette smoking is also a cause of chronic bronchitis, so differences between the levels of exposure in both groups will be small, or possible none. Hence, it is more reasonable to draw controls from the register of patients hospitalized due to various diseases, and bother much less about the same diagnostic procedures.

It has often been postulated that controls should be comparable in every respect to the cases, except for a disease of interest. This way of thinking seems to remind of thoughtless copying the design of experiment, where both control and experimental groups should be comparable, except for the intervention in question. This rule has no practical application in the case control study, just because ideal matching of cases with controls is not possible. Moreover, too precise matching by different variables could easily reduce exposure differences in the compared groups.

The number of control groups in a case-control study has been a matter of argument. Some authors claim that there should be one control group only. Two or more groups would be legitimate if one group is deficient for some reasons. Others believe, that a case-control study consisting of two control groups is the best, because consistent outcomes with both control groups are reinforcing the external validity and entitle for firm extrapolation of results to the target population.

If sufficient number of cases and controls are available, and there are no problems with retrieving information from both cases and controls, then the size of both groups should be the same. This issue becomes a little more complicated when the number of available cases is small or acquiring necessary information is difficult. In such circumstances the ratio of the number of controls per case should be 2:1, 3:1, or even 4:1. Increasing these proportions over 4:1 does not change the statistical power of the study. If more control groups are allowed in the study, it is not necessary to keep them at equal size.

In summary, there are definitely more potential advantages of having hospital patients as control group. First of all, they are easily available for the examinations required by the study, the patients have enough free time, and are more cooperative. Besides, they are in the same "psychological" setting as the case group, they are treated in the same way

by hospital staff, and follow the usual hospital routine. These circumstances minimize the history taking bias, as in the hospital setting patients seem to recollect easier health related hazards from the past. A disadvantage of hospital controls is potentially similar etiologic exposure in both study groups, but to avoid this kind of bias, it is recommended to draw controls from different diagnostic categories.

Assessment of past nutritional exposure

In order to be valid, the assessment of past dietary habits should be performed precisely. Usually, the timing and duration of past dietary exposure may be easier and more precisely assessed than the level of dietary exposure, but combining the duration of exposure with its level helps to estimate an approximated dose of exposure in question if possible. Also the use of biomarkers indicating the accumulated dose of the given dietary factor would be highly desirable.

Past nutritional exposures may be measured in various ways depending on research purposes. The dietary assessment may be based on the foods that people eat, the nutrient or non-nutrient components of foods, biochemical or clinical measures of nutritional status. Since food consists of many substances, food intake is not equivalent of nutrient intake. Although fruits or vegetables are rich sources of vitamins, but it is not proper to think of consumption of fruits or vegetables only in terms of vitamin intake because associations between fruit or vegetable intake and disease risk might be due to other components of fruits or vegetables. We encounter the same problem with regard to non-nutritive components of foods.

On the whole, the assessment of dietary intake is usually based on 1. dietary recalls, 2. food records, 3. dietary histories and 4. food frequency questionnaires. ***Dietary recall*** is simply the full list the foods the respondents report during a short period of time, usually in the preceding 24 or 48 hours. This quick and simple method is appropriate to obtain information on current diet but not diet in the past, since the retrieval of detailed information on past diet is not possible using this kind of tool. Moreover, one single 24-hour recall is not sufficient to characterize individual typical dietary pattern and therefore many recalls must be repeated over longer time period. Nevertheless, a single 24-hour recall collected from a group of individuals can be used to estimate the mean nutrient intakes of this groups but not of individuals.

In ***food record*** study, subjects record the food items they eat for a longer period and it is often combined with information about portion sizes. Sometimes, subjects are also asked to weigh their food portions before consumption or provide a duplicate meal for analysis. Food records study can provide a good representation of typical and current dietary intake if subjects cooperate reasonable well and records are collected over a sufficient period.

In a ***diet history study***, respondents are asked open-ended questions about their usual (present or past) dietary intake. The interviewer inquires about food consumption meal by meal to get the usual pattern of consumption of individuals. This approach can provide detailed picture of individual eating habits, food preparation practices and seasonal variations in food preferences. However, dietary histories are judgmental in a sense that the answers may rather reflect what the subjects think they eat, than what is really eaten.

Food frequency questionnaires like diet histories, focus on usual food intake but the personal interview is precisely structured. Respondents have to answer an interviewer-administered or self-administered questionnaire on frequency of food items listed in the questionnaire. Some types of food frequency questionnaires contain open-ended questions, but mostly the questionnaires use closed-ended questions with predefined response categories. For example, in the open-ended type of questionnaire, respondents are asked how often they eat apples or fish but in the closed-ended type questions, they are asked whether and how often they eat apples daily, or per week. Of course, the quality of the food frequency interview depends not only on the quality of the questionnaire and interviewing techniques, but also on good cooperation with respondents. Food frequency questionnaires provide a reasonable assessment of usual current or past dietary habits and they are commonly used in epidemiologic research in chronic diseases like cardiovascular diseases or cancer.

Food frequency questionnaires may also be used to analyze the intake of nutrients, if one converts the food consumption data into nutrient intakes. This requires the use of high quality food composition tables or nutrient databases. Unfortunately, the quality of food composition data varies from nutrient to nutrient. If purpose of the study is to investigate the impact of foods and not nutrients, then there is no need to consider the potential inadequacies of nutrient databases. Advantage of reporting findings in terms of food consumption lies in avoiding weakly founded assumptions about the exact active food nutrient, and in easy conversion of the findings into practical dietary recommendations. Even though the exact species of fruits that protect against colorectal cancer may not yet been identified, the results of such a study have practical value because they may be used for establishing dietary guidelines.

Having in mind the difficulties in measuring dietary intake, it might seem simpler to assess nutritional status using biomarkers, such as blood or urine nutrient levels. In fact, in some situations, biomarkers may be preferred in the evaluation of exposure to a particular dietary factor. Some biomarkers such as urinary levels of potassium or sodium are considered good indicators of the intakes of these minerals, but others such as blood vitamin A levels are unrelated to dietary vitamin A intake. In addition, the levels of certain nutrients in urine or blood serum may change within hours or days, others change more slowly, reflecting dietary intake over a period of weeks or months. Only few of the currently available biomarkers reflect reasonably well long-term intakes of nutrients over a longer period. Some biomarkers of nutrients are influenced by other factors and in such instances, though the biomarker may still be a good indicator of nutritional status, but it may not accurately reflect dietary intake (for example, smoking reduces blood levels of vitamin C and carotenoids).

Sources of bias in case-control studies

In the case-control studies, investigators search for risk factors retrospectively i.e., after the final health diagnosis. Retrospectively collected data on the exposure may have different sources of bias, for example due to incomplete recalling of prior events by the study subjects (recall bias). Other sources of bias may be simply due to their ignorance of facts related to the past diseases or risk factors. The information about potential causes of

a disease may depend largely on the stage of disease and health condition of the patient, his perception of the cause of his disease, and willingness to participate in the study.

Selection bias of patients to hospital treatment is due to differences in a system of patients' referral to the hospitals by the primary care physicians. This may be a quite arbitrary procedure, and some patients may have a better chance of hospital admission because they have very soon been referred to particular examinations, while others have been deprived of such opportunity. It is extremely difficult to avoid detection bias due to the referral system. Cases of a disease are diagnosed as a result of the complex diagnostic procedures, and it is hardly possible to define all forces governing the system of patient's selection to hospital treatment.

When selection of controls involves sampling from the entire cohort, selection bias is a minor concern, although bias may still occur. When controls are selected from persons with other diseases, considerable care must be taken in specifying the diseases that form the control group. In particular, when a selected disease may not correctly reflect the exposure pattern in the target population. In avoiding this kind of bias, one approach is to include only diseases that are thought to be unrelated to exposure, but this requirement may be difficult to fulfill in practice because adequate evidence for given exposure effects in many diseases is often not available. An alternative approach is to select as controls a sample of all other diseases. This latter approach is generally more reliable because there are few factors that markedly increase risk of various diseases.

Information bias results when the method of data collection makes the information obtained from cases and controls different in a misleading way. For example, cases may recall past events differently than healthy controls do because they are motivated to pay more attention to the causes of disease; this is called recall bias. If the interviewer knows whether subjects are cases or controls (for example, because cases are visibly ill), the conduct of the interview may change in subtle ways, leading to interviewer bias. Epidemiologists who conduct case-control studies need to plan their research so that both recall bias and interviewer bias are reduced as much as possible. Information bias may also occur when biological markers are used as an index of nutritional status. The levels of some markers in the cases may be modified by the onset of disease.

When making assessments of exposure, it is important to allow for the latency period of chronic disease. So far as cases are concerned, it is quite simple to allow for the induction time regarding the last five or ten years of exposure prior to diagnosis. The exposure of the referent subjects should be considered over the same period of time as for the case.

Data analysis

The objectives of data analysis are to determine the direction and strength of association between particular dietary habits and health outcomes. The strength of an association are usually assessed by a relative risk (RR) or an odds ratio (OR). Relative risk is the ratio of the disease rate (incidence) among persons exposed to a given dietary factor divided by the incidence rate among persons not exposed. If the relative risk is greater than one, people exposed to the factor have an increased risk of the outcome under investigation.

If the relative risk is less than one, people exposed to the factor have a decreased risk of the outcome.

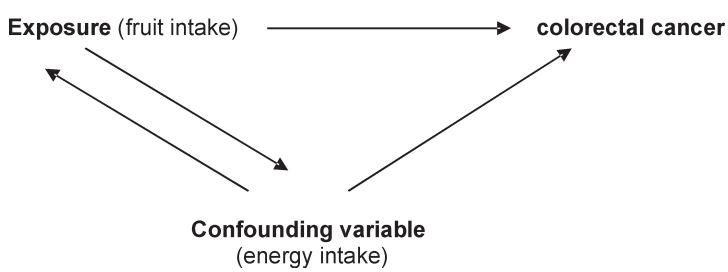
In case-control studies, incidence rates cannot be computed, but one can calculate the ratio of proportions of exposure found in persons with and without disease (Table 5.1). The interpretation of OR is similar as in RR, i.e., if OR is greater than one, people exposed to a given factor run a greater risk for the health outcome. However, it is important to be aware that the rate of disease in the groups is arbitrary and depends on the study design. For example, in a case-control study with one control per case, the rate of disease in the total group is always 0.5 and in the study with 3 controls per case is always 0.25.

Table 5.1. The sequence of steps in case-control study and the logic of the data analysis

1. Step. Choose cases and controls

		Cases	**Controls**
2. Step: assess the past exposure (diet) in cases and controls	Exposure present	A	B
	Exposure absent	C	D
	Total	A + C	B + D
3. Step: compare prevalence of exposure in the groups	A/A + C versus B/B + D		
4. Step: calculate the odds of exposure in cases and controls (OR)	OR = A/C : B/D = A x D/C x B		
5. Step: calculate exposure attributable risk percent	AR% = p (OR – 1)/p(OR –1) + 1		

In order to get valid results from data analysis, it is necessary to consider the possible effects of ***confounding factors*.** By definition, confounder is a variable that is associated not only with the exposure under investigation, but also with the adverse health outcome in question, however it may have no intermediate effect on the analyzed pathologic process. If not accounted for, the confounders may bias the estimated impact of the study factor on health status. Confounders may have both "negative direction" (deflating the impact of exposure) and "positive direction" (inflating this impact). In the extreme cases they may even completely change the direction of associations. Choice of confounders must be very carefully done and one has not to treat a variable as a confounder when it is actually a part of a causal pathway. In a study on diet and colorectal cancer one has to adjust for confounders known to influence colorectal cancer risk, such as age, gender or physical activity. Example below illustrates the effect of confounders on the interpretation of study outcomes.



In dealing with confounders one can analyze data separately for subjects who fall into different categories of the confounding factor. For example, the data may be separately analyzed for men and women, physical activity level or different age groups. Energy intake is also an universal confounder in nutritional epidemiology studies since it is positively correlated with intakes of most nutrients. The interpretation of epidemiologic findings may be misleading if this fact is not taken into account.

In many instance statistical techniques to adjust for the effects of confounding factors are the method of choice. Multivariate statistical analysis is used in situations, where several confounders must be simultaneously accounted for. There are many techniques of multivariate analysis, ranging from simple cross-classification and adjustment to more complex methods of statistical regression analysis. Multivariate techniques help to determine which of the variables have an independent association with the outcome, to detect interactions among variables and to measure the relative contribution of each variable to the risk of the disease.

Advantages of case-control studies. They are relatively quick and inexpensive and they can be applied to common and rare diseases and can investigate a wide variety of potential risk factors simultaneously. Another advantage of this type of study is that it is possible to match subjects for other important characteristics or factors that are not currently under investigation but may confound the results. For example, in a case-control study of peptic ulcer, where the risk is greatly influenced by cigarette smoking, one could select controls with histories of smoking as similar as possible to those of the cases, so that attention could be focused on other factors such as diet. If matching is not performed, it is important to collect information on factors that may influence risk, so that appropriate adjustment for these factors can be made in the data analysis.

Disadvantages of case-control studies. Since this type of investigation requires to collect information about the subjects' past dietary habits this poses a difficult task. People's memories about past dietary habits are imperfect and objective data based on biological markers are usually not available. Case-control studies are also subject to other types of bias, including selection bias and information bias. Selection bias occurs when the cases and controls are selected from different populations or when the subjects in either group are not representative of the population from which they come.